Kaggle：Medical Condition Classification

系所：大數據科技及管理研究所

教授：張詠淳教授

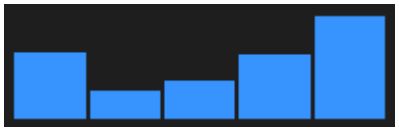
報告者：張哲安

臺北醫學大學

TAIPEI MEDICAL UNIVERSITY

📌 資料

- 📌 相同文本，多個不同標籤
  - 📌 訓練資料4999筆文本相同資料，7995筆完全沒重複
    - ⚠ 多標籤分類
      （預測時，相同文本都使用Logit值最大的）
  - 📌 Kaggle資料中有46筆重複文本，與訓練有669筆重複
    - 📌 雜訊模型無法很好的學習，要分哪一類。
      - ✅ 先切分80%訓練資料，20%驗證資料
        保持驗證資料與Kaggle資料有類似狀況
      - ⚠ 隨機刪除留一筆
      - ⚠ 按照類別比例，留取一筆

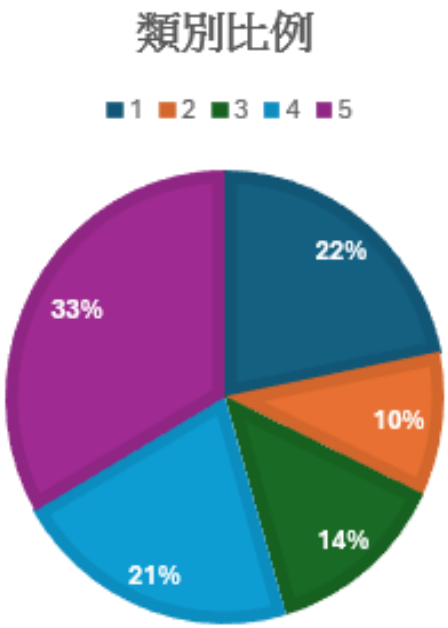


      - ✅ 刪除重複的文本

📌 不平衡資料

📌 數據增強
（只增加少數類樣本，每一筆資料先一次增強，剩下不夠再隨機抽做增強）

- 📌 變為平衡資料
- 📌 增加文本多樣性
  - ⚠ 同義詞替換
  - ⚠ 翻中文在翻回英文
  - ⚠ 上下文詞替換
  - ⚠ Gemini改寫



📌 訓練時加權重

- ✅ Class weight
- ✅ Focal loss

⚠ 訓練總資料量變多，但跟加權重的結果不會差太多

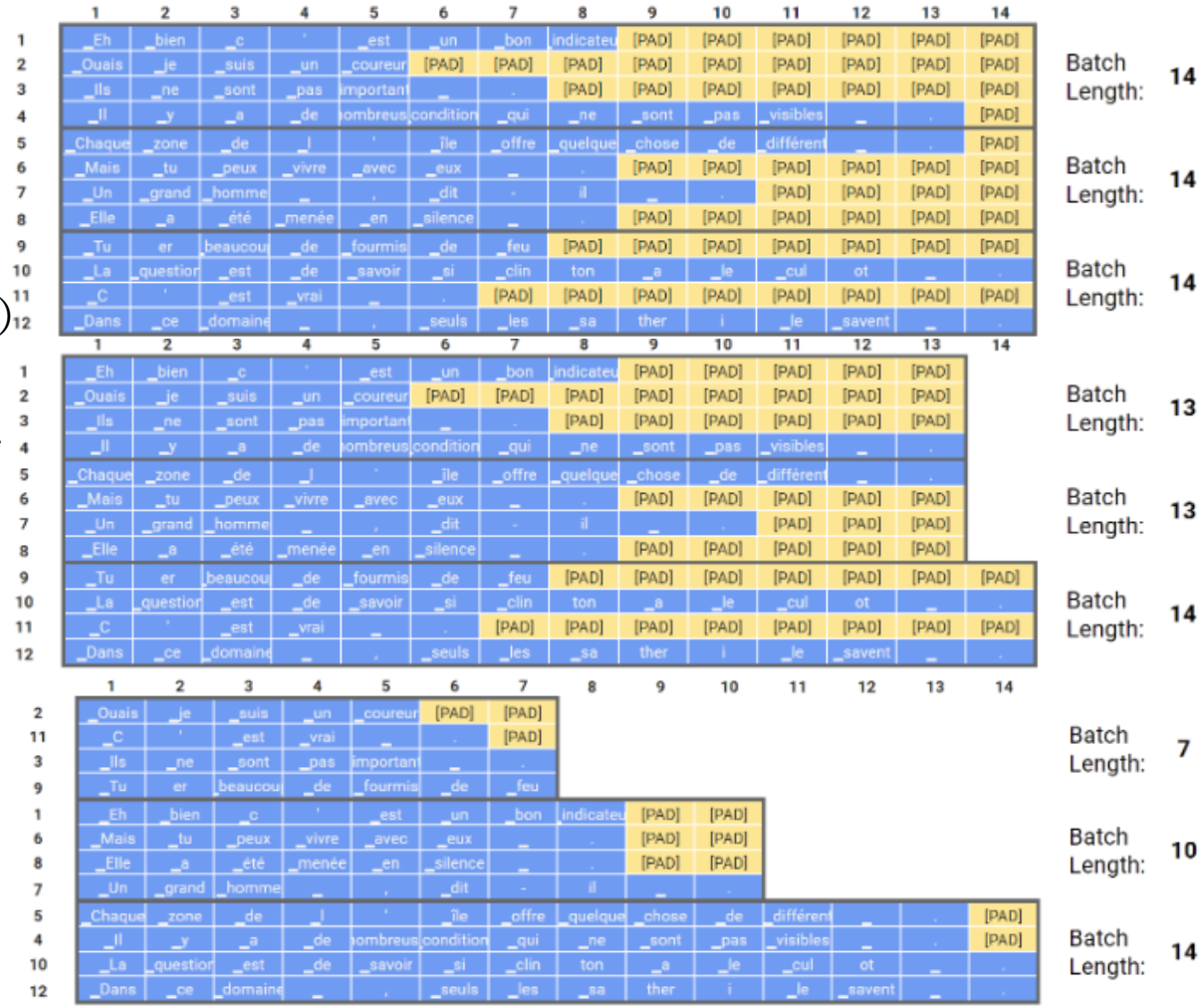
轉Token

- Max length : 512
- Truncation : True
- 動態Padding

✅減少Padding，提升效率(主要)

✅減少冗長雜訊，提升模型表現

✅Kaggle資料也會先進行排序，再回復原本順序



臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

📌 模型

- 📌 Esemble soft voting
  - 📌 Logit
    - ⚠️ Mean
    - ⚠️ Concat
    - ✅ Weighted

    等權重跟單存Voting，會被較差的影響過大
  - 📌 疊加更多BERT

    模型架構變複雜
  - 📌 CLS
    - ⚠️ Mean
    - ⚠️ Concat
    - ⚠️ Weighted

    維度變大(768、1536維)

    ⚠️ 直接分5類(丟失太多資訊)

    後面增加全連階層(捕捉更多特徵資訊)

⚠️ 模型架構越複雜，越難把模型訓練好，Overfitting越嚴重

📌模型

📌BioBert

- ✅Pubmed
- ✅12層

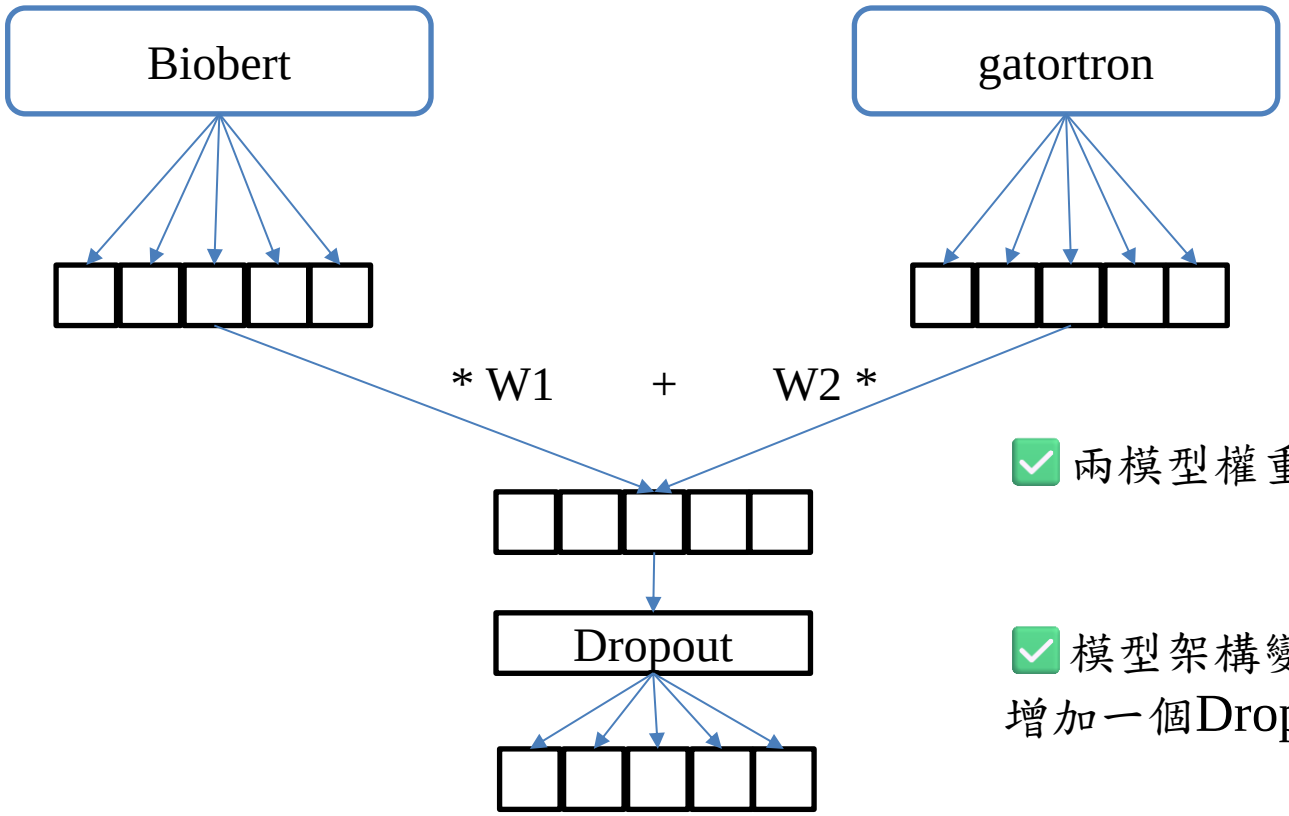
📌Gatortron

- ✅WikiText
- ✅MIMIC-III
- ✅Pubmed
- ✅University of Florida Health System de-identified clinical notes
- ✅24層

✅兩種不同BERT，可以捕捉到不一樣的文本特徵資訊，在最後的CLS或Logit有差異

✅單一皆可以有不差且差不多的模型表現，但在一些分類結果上仍有差異，再進行融合才可以達到互補

📌模型

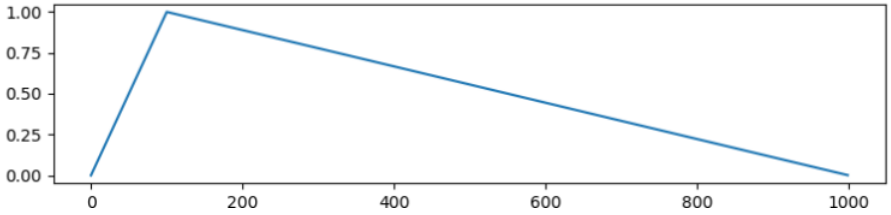


✅兩模型權重，訓練時讓模型自行決定

✅模型架構變複雜，預防Overfitting，增加一個Dropout

📌 參數設定

- ✅ Batch size : 2
- ✅ Gradient accumulation steps : 8
- ✅ Optimizer : adamw torch fused
- ✅ Learning rate : 2e-5
- ✅ Warmup ratio : 0.1
- ✅ Lr scheduler type : linear
- ✅ Epochs : 10
- ✅ Early stopping : 2
- ✅ Metric for best model : f1 macro
- ✅ Fp16 : True
- ✅ Dropout : 0.1



臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

📌 總結

✅ 資料前處理最重要

✅ Domain Fine Tune與任務文本相似>Domain Fine Tune與任務文本較沒關>架構

✅ BioBert > Clinical ModernBert > Neobert

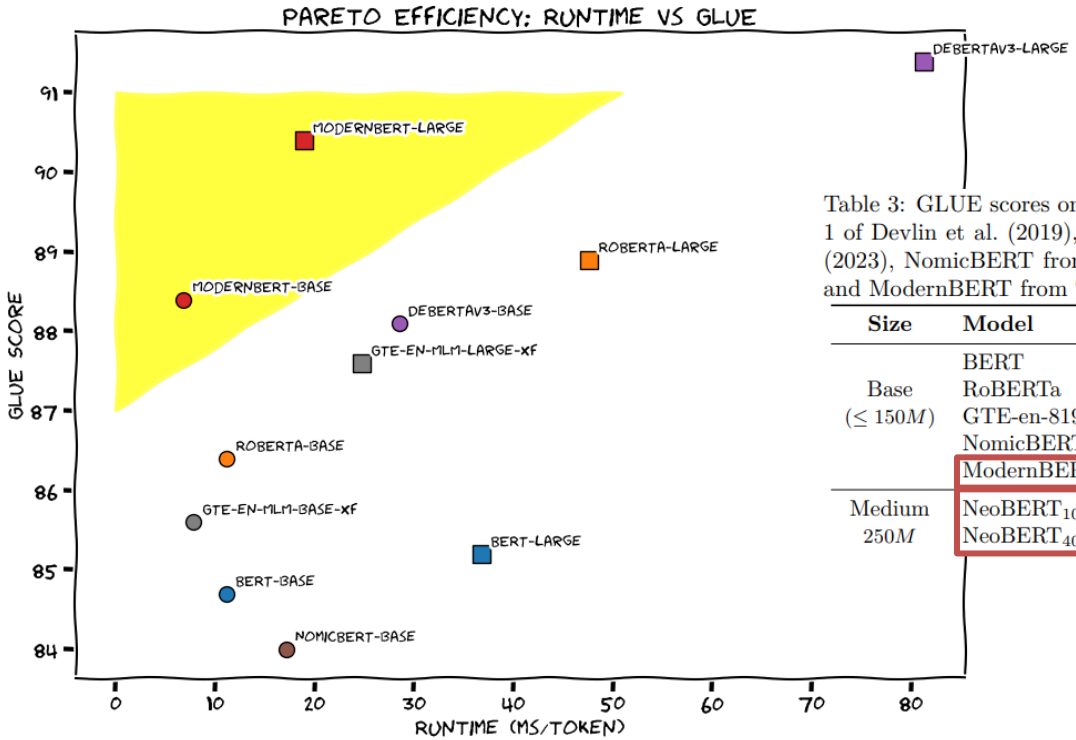


Table 3: GLUE scores on the development set. Baseline scores were retrieved as follows: BERT from Table 1 of Devlin et al. (2019), RoBERTa from Table 8 of Liu et al. (2019), DeBERTa from Table 3 of He et al. (2023), NomicBERT from Table 2 of Nussbaum et al. (2024), GTE from Table 13 of Zhang et al. (2024), and ModernBERT from Table 5 of Warner et al. (2024).

Size	Model	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	Avg.
Base ($\leq 150M$)	BERT	84.0	90.5	71.2	66.4	93.5	88.9	52.1	85.8	79.6
	RoBERTa	87.6	92.8	91.9	78.7	94.8	90.2	63.6	91.2	86.4
	GTE-en-8192	86.7	91.9	88.8	84.8	93.3	92.1	57.0	90.2	85.6
	NomicBERT_{2048}	86.0	92.0	92.0	82.0	93.0	88.0	50.0	90.0	84.0
	ModernBERT	89.1	93.9	92.1	87.4	96.0	92.2	65.1	91.8	88.5
Medium $250M$	NeoBERT_{1024}	88.9	93.9	90.7	91.0	95.8	93.4	64.8	92.1	88.8
	NeoBERT_{4096}	89.0	93.7	90.7	91.3	95.6	93.4	66.2	91.8	89.0

📌 參考資料

[1] C. McCormick, "Smart Batching Tutorial - Speed Up BERT Training," McCormickML, Jul. 29, 2020. [Online]. Available: https://mccormickml.com/2020/07/29/smart-batching-tutorial/

[2] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *Proc. China National Conf. on Chinese Computational Linguistics*, Cham: Springer, 2019, pp. 194–206.

[3] Dataroots, "Fine-tuning BERT for an unbalanced multi-class classification problem," Dataroots Blog. [Online]. Available: https://dataroots.io/blog/incident-team-prediction

[4] E. Alsentzer, J. R. Murphy, W. Boag, W. H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[5] Hugging Face, "Optimizer schedules," Hugging Face Transformers Documentation. [Online]. Available: https://huggingface.co/docs/transformers/v4.52.2/en/main_classes/optimizer_schedules#transformers.SchedulerType

[6] Hugging Face, "Trainer class," Hugging Face Transformers Documentation. [Online]. Available: https://huggingface.co/docs/transformers/main_classes/trainer

[7] Hugging Face, "TrainingArguments," Hugging Face Transformers Documentation. [Online]. Available: https://huggingface.co/docs/transformers/main_classes/trainer#transformers.TrainingArguments

[8] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[9] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020. doi: 10.1093/bioinformatics/btz682.

[10] L. L. Breton, Q. Fournier, M. E. Mezouar, and S. Chandar, "NeoBERT: A Next-Generation BERT," *arXiv preprint arXiv:2502.19587*, 2025.

Thank you